Measuring the Effects of Starting Pitching

Lee Przybylski

Iowa State University

November 18, 2021

Outline

Outline

1. Motivation
2. Generalized Linear Mixed Effects Models
3. Model Selection
4. Predictive Value for Game Outcomes
5. Pitching Metrics
6. Future Work

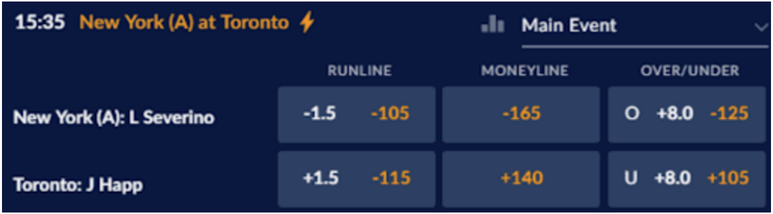
Github: https://github.com/przybylee/RunsScoredAnalysis

Motivation

A sports book presents the moneyline with the team names and the probable starting pitchers.

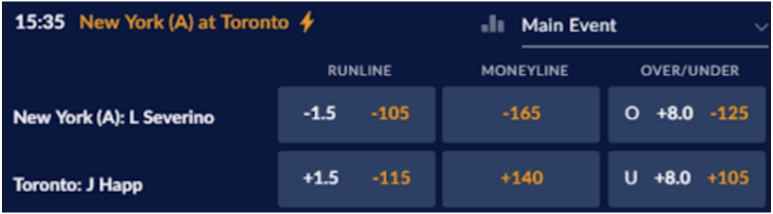
Motivation

A sports book presents the moneyline with the team names and the probable starting pitchers.



Motivation

A sports book presents the moneyline with the team names and the probable starting pitchers.



Can we use this information to make a model that uses this information to predict the outcome of future games?

Can we use predictors from this model to gain a better understanding of teams, venues, and starting pitchers?

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

$$log(\lambda_{ijkl}) = \mu + \omega_i + \delta_j + \nu_k + \chi \mathbf{1}_{ik}.$$

Model Selection

Since runs scored takes values in $\mathbb{N}_0 = \{0, 1, 2, 3, ...\}$, we might consider a model where $y_{ijkl} \overset{\text{iid}}{\sim} \text{poiss}(\lambda_{ijkl})$ is the number of runs scored by team i against team j at venue k during game l, and

$$log(\lambda_{ijkl}) = \mu + \omega_i + \delta_j + \nu_k + \chi \mathbf{1}_{ik}.$$

We let $\mathbf{1}_{ik} = 1$ if team i is playing at home in game k and $\mathbf{1}_{ik} = 0$ otherwise.

Model Selection

Because this model is prone to overdispersion, we choose a generalized linear mixed effects model. This will help our inference.

Model Selection

Because this model is prone to overdispersion, we choose a generalized linear mixed effects model. This will help our inference. We let $y_{ijklm} \sim \text{Poisson}(\lambda_{ijklm})$ be the number of runs scored by team i against team j at venue k, against starting pitcher l during game m. We have

$$\log(\lambda_{ijklm}) = \mu + \chi \mathbf{1}_{im} + b_i + f_j + v_k + p_l + g_m + e_{im}, \tag{1}$$

$$b_i \overset{\text{iid}}{\sim} N(0, \sigma_b^2), f_j \overset{\text{iid}}{\sim} N(0, \sigma_f^2), v_k \overset{\text{iid}}{\sim} N(0, \sigma_v^2), p_l \overset{\text{iid}}{\sim} N(0, \sigma_p^2),$$

$$g_m \overset{\text{iid}}{\sim} N(0, \sigma_g^2), e_{im} \overset{\text{iid}}{\sim} N(0, \sigma_e^2).$$

ERA Leaders

Name	Team	ERA	FIP	WARP	DRA	SPR
Aaron Loup	NYM	0.950	2.440	1.000	3.950	-0.002
Jacob deGrom	NYM	1.080	1.230	3.300	2.410	-0.030
Dominic Leone	SFO	1.510	3.070	0.600	4.500	-0.001
Collin McHugh	TAM	1.550	2.120	1.400	3.590	-0.011
Jesse Chavez	ATL	2.140	2.010	0.500	4.240	-0.006
Tyler Rogers	SFO	2.220	3.280	0.800	4.630	-0.019
Louis Head	TAM	2.310	3.110	0.200	5.000	-0.001
Drew Smith	NYM	2.400	4.690	0.300	4.880	0.008
Corbin Burnes	MIL	2.430	1.630	5.500	2.630	-0.037
Ryan Burr	CWS	2.450	4.230	0.300	4.720	0.001

Variance of OPS

The natural estimators for the parameters determine OPS.

Variance of OPS

The natural estimators for the parameters determine OPS.

$$\text{OBP} = \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_w$$

Variance of OPS

The natural estimators for the parameters determine OPS.

$$\text{OBP} = \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_w$$

$$\text{SLUG} = \frac{\text{PA}(\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4)}{\text{PA}(1 - \hat{p}_w)} = \frac{\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4}{1 - \hat{p}_w - \hat{p}_s}$$

Variance of OPS

The natural estimators for the parameters determine OPS.

$$\text{OBP} = \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \hat{p}_w$$

$$\text{SLUG} = \frac{\text{PA}(\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4)}{\text{PA}(1 - \hat{p}_w)} = \frac{\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4}{1 - \hat{p}_w - \hat{p}_s}$$

$$\text{OPS} = \hat{p}_w + \hat{p}_1 + \hat{p}_2 + \hat{p}_3 + \hat{p}_4 + \frac{\hat{p}_1 + 2\hat{p}_2 + 3\hat{p}_3 + 4\hat{p}_4}{1 - \hat{p}_w - \hat{p}_s}.$$

Delta Method

Using Taylor's theorem, we can show that

$$\mathbb{E}[\mathrm{OPS}] \approx p_w + p_1 + p_2 + p_3 + p_4 + \frac{p_1 + 2p_2 + 3p_3 + p_4}{1 - p_w - p_s} \tag{2}$$

Delta Method

Using Taylor's theorem, we can show that

$$\mathbb{E}[\text{OPS}] \approx p_w + p_1 + p_2 + p_3 + p_4 + \frac{p_1 + 2p_2 + 3p_3 + p_4}{1 - p_w - p_s} \tag{2}$$

and

$$\text{Var}[\text{OPS}] \approx D' \text{Var}[\hat{\boldsymbol{\theta}}] D, \tag{3}$$

Delta Method

Using Taylor's theorem, we can show that

$$\mathbb{E}[\text{OPS}] \approx p_w + p_1 + p_2 + p_3 + p_4 + \frac{p_1 + 2p_2 + 3p_3 + p_4}{1 - p_w - p_s} \tag{2}$$

and

$$\text{Var}[\text{OPS}] \approx D'\text{Var}[\hat{\boldsymbol{\theta}}]D, \tag{3}$$

where

$$D = \begin{bmatrix} \frac{\partial \text{OPS}}{\partial p_s} \\ \frac{\partial OPS}{\partial p_w} \\ \frac{\partial OPS}{\partial p_1} \\ \frac{\partial OPS}{\partial p_2} \\ \frac{\partial OPS}{\partial p_3} \\ \frac{\partial OPS}{\partial p_4} \end{bmatrix} = \begin{bmatrix} \frac{p_1+2p_2+3p_3+4p_4}{(1-p_w-p_s)^2} \\ 1 + \frac{p_1+2p_2+3p_3+4p_4}{(1-p_w-p_s)^2} \\ 1 + \frac{1}{1-p_w-p_s} \\ 1 + \frac{2}{1-p_w-p_s} \\ 1 + \frac{3}{1-p_w-p_s} \\ 1 + \frac{4}{1-p_w-p_s} \end{bmatrix}.$$

Delta Method

We find that

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}] = \mathrm{PA}^{-1}\begin{bmatrix} p_s(1-p_s) & -p_s p_w & -p_s p_1 & -p_s p_2 & -p_s p_3 & -p_s p_4 \\ & p_w(1-p_w) & -p_w p_1 & -p_w p_2 & -p_w p_3 & -p_w p_4 \\ & & p_1(1-p_1) & -p_1 p_2 & -p_1 p_3 & -p_1 p_4 \\ & & & p_2(1-p_2) & -p_2 p_3 & -p_2 p_4 \\ & & & & p_3(1-p_3) & -p_3 p_4 \\ & & & & & p_4(1-p_4) \end{bmatrix}$$

Delta Method

We find that

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}] = \mathrm{PA}^{-1} \begin{bmatrix} p_s(1-p_s) & -p_s p_w & -p_s p_1 & -p_s p_2 & -p_s p_3 & -p_s p_4 \\ & p_w(1-p_w) & -p_w p_1 & -p_w p_2 & -p_w p_3 & -p_w p_4 \\ & & p_1(1-p_1) & -p_1 p_2 & -p_1 p_3 & -p_1 p_4 \\ & & & p_2(1-p_2) & -p_2 p_3 & -p_2 p_4 \\ & & & & p_3(1-p_3) & -p_3 p_4 \\ & & & & & p_4(1-p_4) \end{bmatrix}$$

which implies

$$\mathrm{Var}[\mathrm{OPS}] \approx \frac{\sigma^2}{\mathrm{PA}}, \;\; \sigma^2 > 0.$$

Residuals

To test this theory, we fit a mixed effects model to $n = 72087$ observations in MLB data,

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma_k + p_k + t_{js} + \varepsilon_{ijkl},$$

$$p_k \sim N(0, \sigma_p^2), t_{js} \sim N(\sigma_t^2), \varepsilon_{ijkl} \sim N(0, \sigma_{ijkl}^2)$$

Residuals

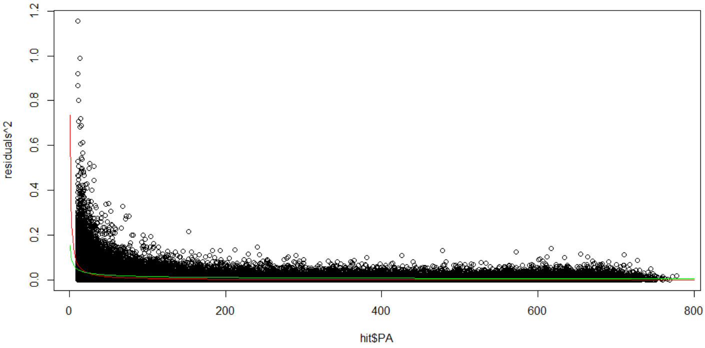


Figure: The red curve represents $e = 0.7375\text{PA}^{-1}$. The green curve represents $e = 0.1532\text{PA}^{-1/2}$.

Residuals

Another approach is to think that if

$$\sigma_{ijkl}^2 = \sigma^2 \mathsf{PA}_{ijkl}^{\theta},$$

Residuals

Another approach is to think that if

$$\sigma^2_{ijkl} = \sigma^2 \mathsf{PA}^{\theta}_{ijkl},$$

or equivalently

$$\log \sigma^2_{ijkl} = \theta \cdot \log \mathsf{PA}_{ijkl} + \log \sigma^2.$$

we should find $\theta = -1$.

Residuals

Another approach is to think that if

$$\sigma^2_{ijkl} = \sigma^2 \mathrm{PA}^{\theta}_{ijkl},$$

or equivalently

$$\log \sigma^2_{ijkl} = \theta \cdot \log \mathrm{PA}_{ijkl} + \log \sigma^2.$$

we should find $\theta = -1$.

Unfortunately, a linear regression using the squared residuals above finds $\hat{\theta} \approx -0.51$.

MLB Models

We test our models on MLB data, and compare our results to the Cramer study. We will use data from every MLB season, to compare each historic major league.

MLBtimeline.png

Birth Year Effects

First we consider if players are improving over time. Consider two models:

Birth Year Effects

First we consider if players are improving over time. Consider two models:

$$OPS_{ijkl} = \mu + \lambda_{ij} + p_k + \varepsilon_{ijkl} \mathrm{PA}_{ijkl}^{-1/2} \tag{4}$$

Birth Year Effects

First we consider if players are improving over time. Consider two models:

$$OPS_{ijkl} = \mu + \lambda_{ij} + p_k + \varepsilon_{ijkl} \mathrm{PA}_{ijkl}^{-1/2} \tag{4}$$

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma x_k + p_k + \varepsilon_{ijkl} \mathrm{PA}_{ijk}^{-1/2} \tag{5}$$

Birth Year Effects

First we consider if players are improving over time. Consider two models:

$$OPS_{ijkl} = \mu + \lambda_{ij} + p_k + \varepsilon_{ijkl}\mathrm{PA}_{ijkl}^{-1/2} \tag{4}$$

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma x_k + p_k + \varepsilon_{ijkl}\mathrm{PA}_{ijk}^{-1/2} \tag{5}$$

$$p_k \sim N(0, \sigma_p^2),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

Birth Year Effects

First we consider if players are improving over time. Consider two models:

$$OPS_{ijkl} = \mu + \lambda_{ij} + p_k + \varepsilon_{ijkl} \mathrm{PA}_{ijkl}^{-1/2} \tag{4}$$

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma x_k + p_k + \varepsilon_{ijkl} \mathrm{PA}_{ijk}^{-1/2} \tag{5}$$

$$p_k \sim N(0, \sigma_p^2),\ \varepsilon_{ijk} \sim N(0, \sigma^2)$$

We fit the models using max. likelihood and find $\hat{\gamma} = 0.0024$, and the likelihood ratio statistic is $\chi_1^2 = 471.43$, which is significant. The average hitting talent of players who reach MLB is increasing each year.

Here are the league effects estimated when assuming $\gamma = 0$. lg.yr_effs_m

League-Year Effects

Here are the league effects estimated by (5) lg.yr_effs_mixed2.p

Final MLB Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma_k + p_k + t_{sj} + \varepsilon_{ijkl}\mathrm{PA}_{ijkl}^{-1/2} \tag{6}$$

$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

Final MLB Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma_k + p_k + t_{sj} + \varepsilon_{ijkl}\text{PA}_{ijkl}^{-1/2} \tag{6}$$

$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
lg.yr	9.29	0.03	279.00	2346.11	15.19	0.0000
birth.yrf	1.44	0.01	161.00	10545.75	4.08	0.0000

Final MLB Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \gamma_k + p_k + t_{sj} + \varepsilon_{ijkl} \mathrm{PA}_{ijkl}^{-1/2} \tag{6}$$

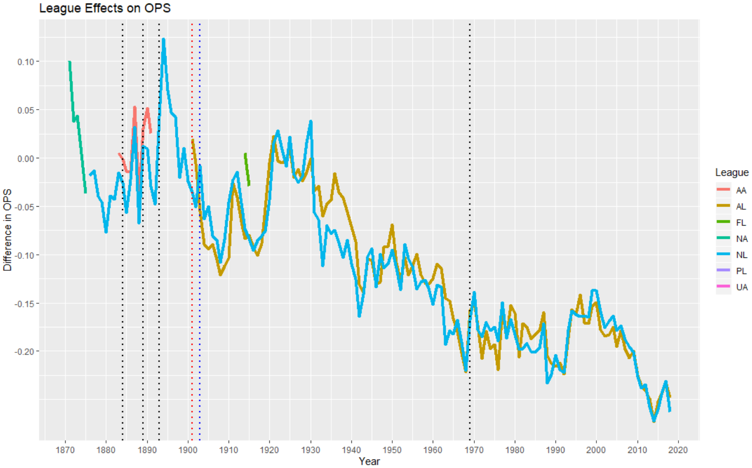
$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
lg.yr	9.29	0.03	279.00	2346.11	15.19	0.0000
birth.yrf	1.44	0.01	161.00	10545.75	4.08	0.0000

$$\hat{\sigma}_p^2 = 0.0166, \hat{\sigma}_t^2 = 0.0003, \hat{\sigma}^2 = 0.0022.$$

MLB Model

Here is a plot of our estimates $\hat{\lambda}_{ij}$.



MLB Model

Here is a plot of our estimates $\hat{\gamma}_k$. MLB_birtheffs.png

Collegiate Data

NCAA			
#Conferences	#Teams	Seasons	n. obs
35	310	2010-2020	57859
Summer			
#Leagues	#Teams	Seasons	
28	409	1996-2019	42648

Collegiate Data

NCAA			
#Conferences	#Teams	Seasons	n. obs
35	310	2010-2020	57859
Summer			
#Leagues	#Teams	Seasons	
28	409	1996-2019	42648

There were 22443 observations in the summer data that were able to be matched to id numbers appearing in the NCAA data.

Cleaning College Data

Year	Team	League	last name	first name
2015	Geneva Red Wings	New York Collegiate	Rodriguez	Alex
2017	Valley Blue Sox	New England Collegiate League	Rodriguez	Alex
2017	Texarkana Twins	Texas Collegiate League	Rodriguez	Alex
2017	Concord Athletics	Southern Collegiate League	Rodriguez	Alex
2017	Riverside Bulldogs	Southern California League	Rodriguez	Alex
2018	North Adams SteepleCats	New England Collegiate League	Rodriguez	Alex
2018	Savannah Bananas	Coastal Plain League	Rodriguez	Alex
2018	Academy Barons	California Collegiate League	Rodriguez	Alex

Final College Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \alpha_{0k} + \alpha_{1k}x_j + \alpha_{2k}x_j^2 + p_k + t_{sj} + \varepsilon_{ijkl}PA_{ijkl}^{-1/2} \tag{7}$$

$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

Final College Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \alpha_{0k} + \alpha_{1k}x_j + \alpha_{2k}x_j^2 + p_k + t_{sj} + \varepsilon_{ijkl}PA_{ijkl}^{-1/2} \qquad (7)$$

$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
lg.yr	36.87	0.07	549.00	4541.19	14.55	0.0000
r.lg	2.32	0.07	34.00	33621.25	14.81	0.0000
r.ssn.sc	0.13	0.13	1.00	12267.65	27.34	0.0000
r.lg:r.ssn.sc	0.23	0.01	34.00	45134.50	1.48	0.0347
r.lg:$(\text{r.ssn.sc})^2$	0.31	0.01	35.00	39059.73	1.93	0.0008

Final College Model

$$OPS_{ijkl} = \mu + \lambda_{ij} + \alpha_{0k} + \alpha_{1k} x_j + \alpha_{2k} x_j^2 + p_k + t_{sj} + \varepsilon_{ijkl} PA_{ijkl}^{-1/2} \tag{7}$$

$$p_k \sim N(0, \sigma_p^2),\ t_{sj} \sim N(0, \sigma_t^2),\ \varepsilon_{ijkl} \sim N(0, \sigma^2)$$

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
lg.yr	36.87	0.07	549.00	4541.19	14.55	0.0000
r.lg	2.32	0.07	34.00	33621.25	14.81	0.0000
r.ssn.sc	0.13	0.13	1.00	12267.65	27.34	0.0000
r.lg:r.ssn.sc	0.23	0.01	34.00	45134.50	1.48	0.0347
r.lg:(r.ssn.sc)^2	0.31	0.01	35.00	39059.73	1.93	0.0008

$$\hat{\sigma}_p^2 = 0.0074, \hat{\sigma}_t^2 = 0.0010, \hat{\sigma}^2 = 0.0046.$$

College Model

Here is a plot of our estimates $\hat{\lambda}_{ij}$ for power 5. college_lgeffs_pow

College Model

Our model provides a way to perform null hypothesis testing.

College Model

Our model provides a way to perform null hypothesis testing.

$$H_0 : \frac{1}{5} \sum_{i \in \mathrm{Pwr5}} (\lambda_{i,2018} - \lambda_{i,2014}) = 0$$

College Model

Our model provides a way to perform null hypothesis testing.

$$H_0 : \frac{1}{5} \sum_{i \in \text{Pwr5}} (\lambda_{i,2018} - \lambda_{i,2014}) = 0$$

We reject H_0:

Estimate	Std. Error	df	t value	lower	upper	Pr(>|t|)
0.96	0.04	4329.49	22.15	0.87	1.04	0.00

MAC vs. Big10

One question we consider is how do hitters that started in the MAC in 2014 compare to their Big10 counterparts, and how do both of these hitters compare to non-DI hitters.

MAC vs. Big10

One question we consider is how do hitters that started in the MAC in 2014 compare to their Big10 counterparts, and how do both of these hitters compare to non-DI hitters.

We use the Satterthwaite method to run the appropriate t-test:

	Estimate	Std. Error	df	t value	lower	upper	Pr(>|t|)
1	0.0320	0.0111	35935.8453	2.8986	0.0104	0.0537	0.0038
2	0.0426	0.0081	35742.6161	5.2768	0.0268	0.0584	0.0000
3	0.0105	0.0087	42439.0905	1.2071	-0.0066	0.0276	0.2274

MAC vs. Big10

One question we consider is how do hitters that started in the MAC in 2014 compare to their Big10 counterparts, and how do both of these hitters compare to non-DI hitters.

We use the Satterthwaite method to run the appropriate t-test:

	Estimate	Std. Error	df	t value	lower	upper	Pr(>|t|)
1	0.0320	0.0111	35935.8453	2.8986	0.0104	0.0537	0.0038
2	0.0426	0.0081	35742.6161	5.2768	0.0268	0.0584	0.0000
3	0.0105	0.0087	42439.0905	1.2071	-0.0066	0.0276	0.2274

1. $\alpha_{0,\text{Big10}} - \alpha_{0,\text{MAC}}$
2. $\alpha_{0,\text{Big10}} - \alpha_{0,\text{nDI}}$
3. $\alpha_{0,\text{MAC}} - \alpha_{0,\text{nDI}}$

Draft Results

Compare those results to the rates at which those hitters were drafted.

Conference	Total Drft. Rate	2014 Drft.Rate
Big10	23.46%	35.29%
MAC	10.57%	10.26%
non-DI	2.93%	5.45%

Future Work and Closing Thoughts

- Our model was pretty consistent with the Cramer study.

Future Work and Closing Thoughts

- Our model was pretty consistent with the Cramer study.
- It did not seem correct to assume that consecutive at bats are independent when modeling OPS.

Future Work and Closing Thoughts

- Our model was pretty consistent with the Cramer study.
- It did not seem correct to assume that consecutive at bats are independent when modeling OPS.
- Cleaning college data was challenging. I'd like to model transitions: $\mathbb{P}[L_{j+1} = l | L_j]$ or $\mathbb{P}[(L, \mathrm{OPS})_{j+1} = (l, x) | (L, \mathrm{OPS})_j]$

Future Work and Closing Thoughts

- Our model was pretty consistent with the Cramer study.
- It did not seem correct to assume that consecutive at bats are independent when modeling OPS.
- Cleaning college data was challenging. I'd like to model transitions: $\mathbb{P}[L_{j+1} = l | L_j]$ or $\mathbb{P}[(L, \mathrm{OPS})_{j+1} = (l, x) | (L, \mathrm{OPS})_j]$
- Need a way to control for age in the MLB OPS model.

Future Work and Closing Thoughts

- Our model was pretty consistent with the Cramer study.
- It did not seem correct to assume that consecutive at bats are independent when modeling OPS.
- Cleaning college data was challenging. I'd like to model transitions: $\mathbb{P}[L_{j+1} = l | L_j]$ or $\mathbb{P}[(L, \mathrm{OPS})_{j+1} = (l, x) | (L, \mathrm{OPS})_j]$
- Need a way to control for age in the MLB OPS model.
- I would like to compare BLUPS for college hitters to their draft results.

Thank You

[1] "The Baseball Cube- MLB, Minor League, College Statistics, Data and the Draft." Accessed October 2019. http://thebaseballcube.com/.

[2] Richard Cramer. "Average Batting Skill Through Major League History." *Baseball Research Journal.* (1980), Retrieved October 2019 from https://ourgame.mlblogs.com/average-batting-skill-through-major-league-history-landmarks-of-sabermetrics-part-i-bb5849adae0b.

[3] John Thorn and Pete Palmer. *The Hidden Game of Baseball*, The University of Chicago Press, 1984.
https://www.si.com/betting/2020/07/02/gambling-101-major-league-baseball-betting